Patrick Quaid
Regression
Fall 2013
Student Project

Controlling Blood Pressure

Introduction

I have been keeping track of my weight and blood pressure, at the request of my wife, for well over a year. I decided that I would like to, based on my data set, determine if it has been more effective for me to reduce my blood pressure by losing weight, or by getting more aerobic exercise. Of course, without a controlled experiment, I cannot draw this sort of cause-effect relationship. However, a strong correlation and good model are a very good starting point from which other experiments and studies can be designed.

Hypothesis and Model Development

I have well over a year of data on my weight and blood pressure (in appendix, as well as statistics workbook). Unfortunately I have not kept an accurate record of how much aerobic exercise I have performed – however, I do have a good idea of what periods of time I have been active in martial arts, which is by far my biggest source of aerobic exercise. Therefore, I will model aerobic exercise using a dummy variable, that takes on a value of 1 when I was active in martial arts, and 0 otherwise.

In order to simplify things, I will focus only on the diastolic reading of my blood pressure (for the rest of the report, Blood Pressure will actually be referring only to the diastolic measurement).

The models I will test are:

i) $Y = \alpha + \beta_1 X_1 + \varepsilon$

ii) $Y = \alpha + \gamma_1 D_1 + \varepsilon$

iii) $Y = \alpha + \beta_1 X_1 + \gamma_1 D_1 + \varepsilon$

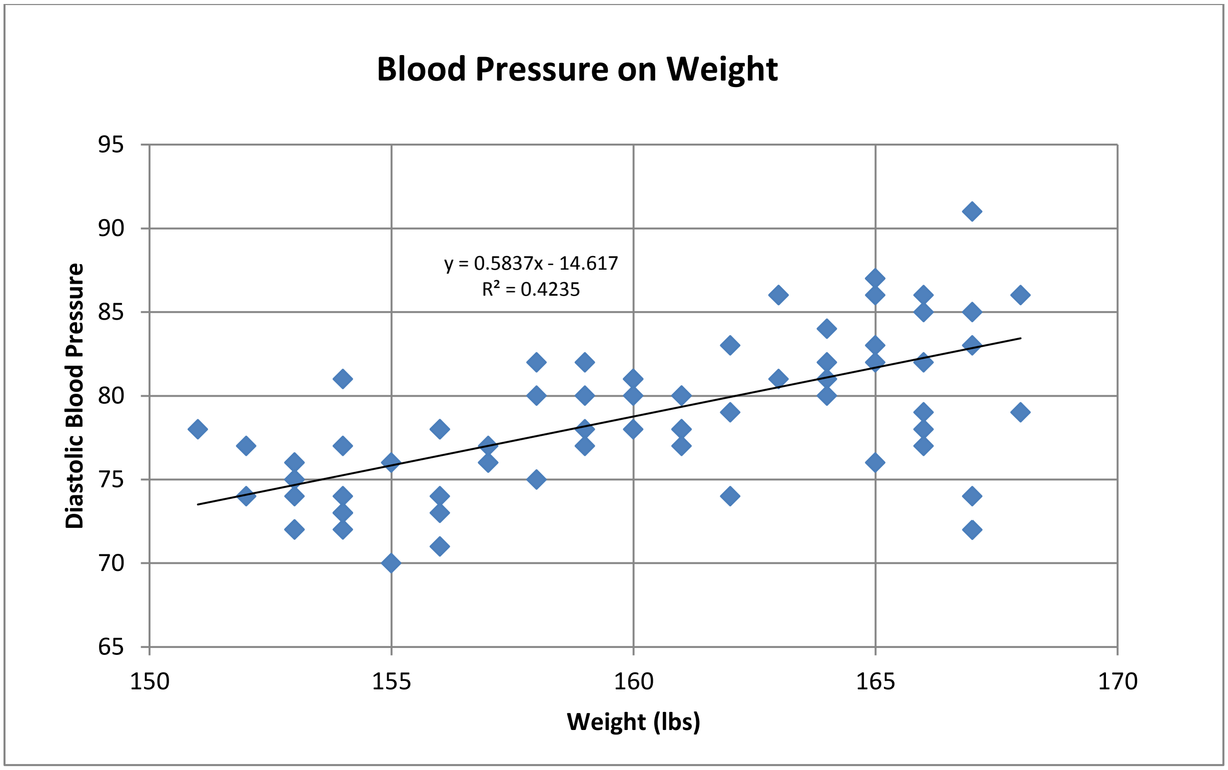
iv) $Y = \alpha + \beta_1 X_1 + \gamma_1 D_1 + \delta_1 (X_1 D_1) + \varepsilon$

where: Y = Blood Pressure
X_1 = Weight
D_1 = 1 for active in martial arts
D_1 = 0 for not active in martial arts

I expect that my weight and whether or not I am active in martial arts will both be strongly correlated to my blood pressure. In particular I expect models (iii) or (iv) to be the best.

Model (i): Y = α + βX + ϵ

I first did a regression of Blood Pressure on Weight. Here is a scatterplot of the data:



From this scatterplot we see a larger variance in Blood Pressure for the largest values of weight. However, there is not a clear pattern of larger values of weight being associated in general with larger variance of blood pressure – note the tight groups of points in the middle of the scatterplot.

Here are 5-number summaries for weight and blood pressure:

Weight		**Blood Pressure**	
Min:	151	Min:	70
H_L:	155	H_L:	75
Med:	160	Med:	78
H_U:	164.5	H_U:	81.5
Max:	168	Max:	91

For weight we have $\frac{H_U - M}{M - H_L} = 0.9$, and for blood pressure $\frac{H_U - M}{M - H_L} = 1.167$. Since these are both close to 1 we have relatively symmetric data sets. And since the minimums and maximums of each data set are within 1.5*Hinge Spread of the upper and lower hinges, none of these observations are considered outliers within their data sets.

Below is a plot of the residuals:

This is a desirable residual plot as there does not appear to be any distinct pattern in the residuals. The random scatter of residuals around the regression line suggests that the linear relationship used is appropriate.

Based on the above statistics and observations, I do not see cause here to transform either set of data, or remove any suspect observations.

The following regression statistics were found using Microsoft Excel's Analysis ToolPak for Regression with Diastolic Blood Pressure as the response variable and Weight as the explanatory variable:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.6508
R Square	0.4235
Adjusted R Square	0.4156
Standard Error	3.4611
Observations	75

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	642.5030	642.5030	53.6348	2.60547E-10
Residual	73	874.4837	11.9792		
Total	74	1516.9867			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-14.6170	12.7419	-1.1472	0.2551	-40.0116	10.7776
Weight	0.5837	0.0797	7.3236	2.60547E-10	0.4248	0.7425

From this output we see our model is Y = -14.617 + 0.5837X.
The null hypothesis Ho: β = 0 has a t-statistic of 7.3236 with 73 degrees of freedom, and corresponding p-value of 2.6055E-10. This extremely low p-value leads us to reject this null hypothesis and conclude that β ≠ 0.
This is consistent with our 95% confidence interval for the slope (0.4248, 0.7425).
Although we have strong evidence of a relationship between weight and blood pressure, the R^2 value = 0.4235, which is a relatively low value (i.e. only 42% of the variation in blood pressure is explained by this regression on weight).
Thus, we have a statistically significant model, but not necessarily a very useful model.
Since we have only regressed on weight we can interpret the estimate of β = 0.5837 to mean that, on average, we see an increase of 0.5837 in the diastolic reading of blood pressure for every 1 lb increase in weight.

Model (ii): Y = α + γD + ϵ
This model only investigates the relationship between martial arts activity (my approach to get at aerobic exercise) and blood pressure. Since the only explanatory variable is categorical it is useful to compare boxplots for the two groups:

	Martial Arts Participation	
Statistics of B.P.	Y	N
min	70	74
q1	74	79.5
Median	76	81
q3	78	83.5
max	87	91

```
                                          _________
            MA = N   --------------------|____|_______|---------------------------

                              __________
MA = Y         --------------|______|______|------------------------------

-----------------------------------------------------------------------------------------------------Blood Pressure
65             70             75             80             85             90
```

From the boxplots we can see that blood pressure is distributed similarly among the two sets of data, i.e., periods when I was participating in martial arts and periods when I was not. However, there is almost a uniform shift upward of about 4 points in the diastolic reading for each of the 5 statistics from the boxplots. This observation is supported by the output from the regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.54031981
R Square	0.29194549
Adjusted R Square	0.28224612
Standard Error	3.83586042
Observations	75

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	442.8774	442.8774	30.0994	5.6298E-07
Residual	73	1074.1092	14.7138		
Total	74	1516.9867			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	81.5484	0.6889	118.3678	3.57E-85	80.1753	82.9214
Martial Arts	-4.9348	0.8995	-5.4863	5.63E-07	-6.7274	-3.1421

This gives a model of: $Y = 81.5484 – 4.9348D$ where D = 1 if active in martial arts, and 0 if not active.

Here we have a t-statistic of -5.486 with 73 degrees of freedom, with an extremely low p-value of 5.63E-07 for the null hypothesis $\gamma = 0$, and so we reject the null hypothesis and conclude that it is statistically significant, at any reasonable significance level, that $\gamma \neq 0$.

This coefficient has a different meaning than β for the last model since our explanatory variable is a dummy variable being used for a categorical variable. What we can conclude here is that when D is 0 (no martial arts activity during that week), the average diastolic blood pressure reading is 81.55. And that readings for weeks where I did participate in martial arts were on average 4.935 points lower, and thus had an average diastolic reading of 76.6.

The 95% confidence interval for γ, the average difference in blood pressure between weeks with and without martial arts, is (-6.7274, -3.1421). Since 0 is not in this range, we can be 95% confident that martial arts activity is associated with lower blood pressure readings.

However, despite these strong findings that martial arts activity is significant, it does not by itself explain well the large variations we see in blood pressure readings, as is evident by the very low R^2 value of 0.2919.

Model (iii): Y = α + βX + γD + ϵ

This model uses both weight and martial arts activity as explanatory variables. Here are the results of the regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.688205
R Square	0.473627
Adjusted R Square	0.459005
Standard Error	3.330210
Observations	75

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	718.485297	359.242649	32.392519	9.26E-11
Residual	72	798.501369	11.090297		
Total	74	1516.986667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	7.442897	14.877413	0.500282	0.618403	-22.214685	37.100478
Weight	0.454545	0.091181	4.985104	0.000004	0.272780	0.636310
Martial Arts	-2.430421	0.928532	-2.617488	0.010790	-4.281416	-0.579426

Based on these regression statistics our model is: Y = 7.4429 + 0.4545X – 2.4304D

This model has an F-statistic: F = 32.39 with 2 and 72 degrees of freedom, with a corresponding p-value of 9.26E-11. Thus we can reject the null hypothesis that β = γ = 0. When looking at the coefficients on the individual explanatory variables, we see that β has a 95% confidence interval of (0.273, 0.636), and γ has a 95% confidence interval of (-4.281, -0.579). These results are consistent with what we saw from the previous two models that regressed each individually. Here we can conclude that β is positive and so each additional pound of weight lost on average will reduce blood pressure, and in addition γ is negative, so remaining active in martial arts will on average yield lower blood pressure readings than when not participating in martial arts. Specifically this model suggests that each lb of weight lost on average corresponds to a 0.4545 decrease in blood pressure, and on average participation in martial arts corresponds to a decrease in blood pressure an additional 2.43 points.

We still, however, have a generally low R^2 value of 0.4736, so there is a lot of variation in the blood pressure readings that remain unexplained by this model.

Model (iv): $Y = \alpha + \beta X + \gamma D + \delta XD + \epsilon$

This is the 'full' model, since it includes both weight and martial arts activity as explanatory variables, and also includes their interaction term XD. Here are the results of the regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.6922
R Square	0.4791
Adjusted R Square	0.4571
Standard Error	3.3361
Observations	75

ANOVA

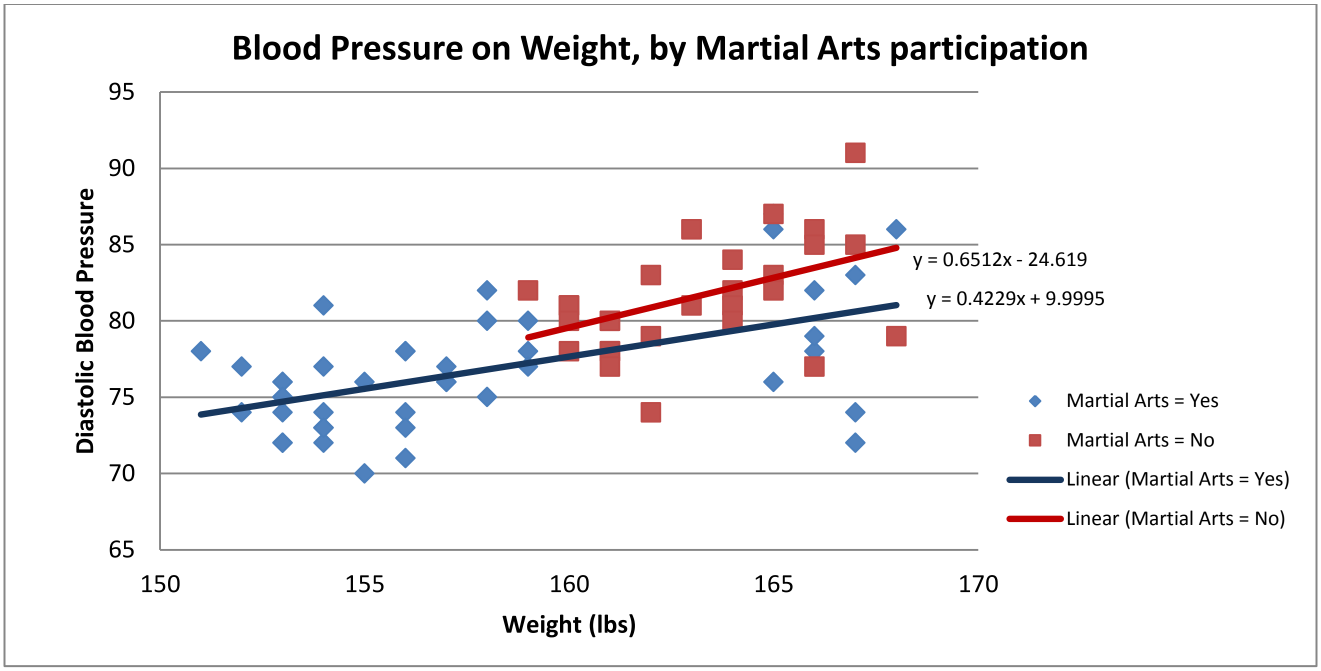
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	726.7904	242.2635	21.767	4.20144E-10
Residual	71	790.1962	11.1295		
Total	74	1516.9867			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-24.6188	39.9956	-0.6155	0.5402	-104.3678	55.1303
Weight	0.6512	0.2453	2.6548	0.0098	0.1621	1.1403
Martial Arts	34.6183	42.8982	0.8070	0.4224	-50.9184	120.1549
Weight x Martial Arts	-0.2283	0.2643	-0.8638	0.3906	-0.7553	0.2987

This regression produces the model: $Y = -24.6188 + 0.6512X + 34.6183D – 0.2283(XD)$

Here again we have a statistically significant model. This model has an F-statistic: $F = 21.767$ with 3 and 71 degrees of freedom, with a corresponding p-value of 4.2E-10. Thus we can reject the null hypothesis that $\beta = \gamma = \delta = 0$. This model assumes that the data sets for martial arts and non-martial arts come from populations with different slopes and intercepts.

What this model does is essentially performs two completely separate regressions on the observations corresponding to martial arts activity, and the points where there was no martial arts activity. We can see that when $D = 1$ the model is $Y = 9.9995 + 0.4229X$, and when $D = 0$ the model becomes $Y = -24.6188 + 0.6512X$. Below is a scatterplot with the individual regression equations plotted:



We can see from the scatterplot that there is a clear positive relationship between weight and blood pressure for both sets of data, and that the non-martial arts points in general are associated with higher levels of blood pressure.

It is tempting to take the full model since it has every possible explanatory variable being considered, and the highest R^2 at 0.4791. However, adding extra explanatory variables will always increase the R^2 value, even if the variables are not actually related to the response variable. In order to compare this model to the previous we can look at the R^2-adjusted which takes into account the degrees of freedom of each model. We can see that the R^2-adjusted actually decreased from 0.459 for the model $Y = \alpha + \beta X + \gamma D$ with no interaction term, to 0.457 for the full model. Since this model is more complicated than the previous, and in addition has a lower R^2-adjusted value, I would not use this model.

Conclusion

From my data set I have found diastolic blood pressure to be positively correlated to weight, and have also found that martial arts participation is associated with decreased blood pressure readings. Of the four models analyzed, the best model appears to be $Y = \alpha + \beta X + \gamma D$ which takes into account both explanatory variables but does not include an interaction term. This model has the highest R^2 adjusted of the four models.

Although I have chosen a 'best' of the models, it is still not a great model, since it's R^2 value is low at only 0.4736. This means that more than 50% of the variation in diastolic blood pressure remains unexplained. For future studies I would suggest tracking and taking into account other possible explanatory variables, such as sodium intake. Also, I would like to measure minutes of moderate to intense aerobic exercise and use that as a quantitative explanatory variable instead of the martial arts participation used in these models.

Appendix

Data: collected by myself every Saturday between July 7 2012 and December 14 2013, using a standard scale, and a digital blood pressure cuff. Martial Arts = 1 if I was actively participating in martial arts during that time period.

Date	Diastolic	Weight	Martial Arts?
7/7/2012	72	154	1
7/14/2012	74	153	1
7/21/2012	75	153	1
7/28/2012	81	154	1
8/4/2012	76	153	1
8/11/2012	74	152	1
8/18/2012	73	154	1
8/25/2012	76	153	1
9/1/2012	72	153	1
9/8/2012	77	152	1
9/15/2012	72	153	1
9/22/2012	75	153	1
9/29/2012	73	154	1
10/6/2012	78	151	1
10/13/2012	74	154	1
10/20/2012	73	154	1
10/27/2012	75	153	1
11/3/2012	70	155	1
11/10/2012	77	154	1
11/17/2012	76	155	1
11/24/2012	74	156	1
12/1/2012	73	156	1
12/8/2012	76	157	1
12/15/2012	78	156	1
12/22/2012	71	156	1
12/29/2012	75	158	1
1/5/2013	78	156	1
1/12/2013	76	157	1
1/19/2013	77	157	1
1/26/2013	82	158	1
2/2/2013	78	159	1
2/9/2013	80	159	1
2/16/2013	77	159	1
2/23/2013	80	158	1
3/2/2013	78	160	0
3/9/2013	81	160	0

3/16/2013	82	159	0
3/23/2013	81	160	0
3/30/2013	74	162	0
4/6/2013	81	163	0
4/13/2013	87	165	0
4/20/2013	86	163	0
4/27/2013	80	161	0
5/4/2013	83	162	0
5/11/2013	80	161	0
5/18/2013	78	161	0
5/25/2013	81	160	0
6/1/2013	78	161	0
6/8/2013	77	161	0
6/15/2013	80	160	0
6/22/2013	79	162	0
6/29/2013	84	164	0
7/6/2013	81	164	0
7/13/2013	86	163	0
7/20/2013	82	164	0
7/27/2013	81	164	0
8/3/2013	83	165	0
8/10/2013	80	164	0
8/17/2013	78	166	1
8/24/2013	74	167	1
8/31/2013	72	167	1
9/7/2013	86	168	1
9/14/2013	82	166	1
9/21/2013	76	165	1
9/28/2013	86	165	1
10/5/2013	79	166	1
10/19/2013	87	165	1
10/26/2013	83	167	1
11/2/2013	77	166	0
11/9/2013	86	166	0
11/16/2013	85	167	0
11/23/2013	85	166	0
11/30/2013	91	167	0
12/7/2013	82	165	0
12/14/2013	79	168	0